Received: 26 July 2019 | Revised: 6 April 2020 | Accepted: 21 May 2020

DOI: 10.1111/geb.13149

RESEARCH PAPER

Global Ecology and Biogeography A Journal of Macroecology WILEY

The strengths and weaknesses of species distribution models in biome delimitation

Peter W. Moonlight[1] | Pedro Luiz Silva de Miranda[2,3] | Domingos Cardoso[4] | Kyle G. Dexter[1,2] | Ary T. Oliveira-Filho[5] | R. Toby Pennington[1,6] | Gustavo Ramos[1,3,7] | Tiina E. Särkinen[1]

[1]Tropical Diversity Section, Royal Botanic Garden Edinburgh, Edinburgh, UK

[2]Department of Geosciences, University of Edinburgh, Edinburgh, UK

[3]Département GxABT, Laboratoire de Foresterie des régions tropicales et subtropicales, Université de Liège, Gembloux, Belgique

[4]National Institute of Science and Technology in Interdisciplinary and Transdisciplinary Studies in Ecology and Evolution (INCT IN-TREE), Instituto de Biologia, Universidade Federal da Bahia, Salvador, Bahia, Brazil

[5]Departamento de Botânica, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[6]Department of Geography, University of Exeter, Exeter, UK

[7]Department of Biological Sciences, University of Edinburgh, Edinburgh, UK

Correspondence
Peter W. Moonlight, Tropical Diversity Section, Royal Botanic Gardens Edinburgh, 20A Inverleith Row, Edinburgh EH3 5LR, UK.
Email: pmoonlight@rbge.org.uk

Funding information
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Grant/Award Number: 99999.013197/2013-04; Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: 308244/2018-4; Royal Society, Grant/Award Number: NAF/R1/180331; Natural Environment Research Council, Grant/Award Number: NE/N012526/1; CNPq

Editor: Antoine Guisan

Abstract

Aim: The aim was to test whether species distribution models (SDMs) can reproduce major macroecological patterns in a species-rich, tropical region and provide recommendations for using SDMs in areas with sparse biotic inventory data.

Location: North-east Brazil, including Minas Gerais.

Time period: Present.

Major taxa studied: Flowering plants.

Methods: Species composition estimates derived from stacked SDMs (s-SDMs) were compared with data from 1,506 inventories of 933 woody plant species from north-east Brazil. Both datasets were used in hierarchical clustering analyses to delimit floristic units that correspond to biomes. The ability of s-SDMs to predict the identity, functional composition and floristic composition of biomes was compared across geographical and environmental space.

Results: The s-SDMs and inventory data both resolved four major biomes that largely corresponded in terms of their distribution, floristics and function. The s-SDMs proved excellent at identifying broad-scale biomes and their function, but misassigned many individual sites in complex savanna–forest mosaics.

Main conclusions: Our results show that s-SDMs have a unique role to play in describing macroecological patterns in areas lacking inventory data and for poorly known taxa. s-SDMs accurately predict floristic and functional macroecological patterns but struggle in areas where non-climatic factors, such as fire or soil, play key roles in governing distributions.

KEYWORDS

biome delimitation, cluster analysis, diversity patterns, macroecology, savanna, species distribution modelling

1 | INTRODUCTION

Macroecology is the study of large-scale patterns of biological diversity across space and time and the underlying community assembly processes that determine these patterns. The description of macroecological patterns is of paramount importance in addressing the global, societal need for planet-wide ecosystem models, which have the potential to transform our understanding of the biosphere (Purves et al., 2013; Socolar, Gilroy, Kunin, & Edwards, 2016). As a science, macroecology relies upon accurate and comprehensive data of species distributions through time and space. Distribution data for mammals, birds and amphibians are becoming increasingly available (Castro-Insua, Gomez-Rodrıguez, & Baselga, 2016; McKnight et al., 2007; Melo, Rangel, & Diniz-Filho, 2009), but distribution data for most lineages in the tree of life are still poor (Scheffers, Joppa, Pimm, & Laurance, 2012). This is known as the Wallacean shortfall.

A suite of methods has been developed to extrapolate our limited knowledge of species distributions to complete mapping of biodiversity patterns. These methods fall into two categories: stacked species distribution models (s-SDMs) and macroecological models (MEMs). Macroecological models relate emergent ecosystem properties, such as species richness or functional diversity, to environmental variables statistically (Gould, 2000). A significant limitation of macroecological models is, however, that they cannot predict species identity; hence, they are used primarily for exploring macroecological phenomena relying on emergent ecosystem properties, such as species richness. s-SDMs, on the contrary, permit the exploration of diversity patterns whose investigation relies upon species identity.

If s-SDMs are to be used in macroecological studies with confidence, their capacity to predict patterns should be evaluated critically to gain a better understanding of their strengths and weaknesses. Testing s-SDMs against known macroecological patterns has focused primarily upon reproducing species richness patterns. s-SDMs typically over-predict species richness compared with both MEMs (Dubuis et al., 2011) and independent data (Cooper & Soberón, 2017; Pineda & Lobo, 2009; Pottier et al., 2013), although this pattern is not universal (D'Amen, Pradervand, & Guisan, 2015). The general overestimation of species richness in s-SDMs has been attributed to individual species distribution models (SDMs) being unable to account for the full spectrum of community assembly processes that limit species distributions, including biotic interactions, dispersal limitation and ecological carrying capacity (Dubuis et al., 2011; Guisan & Rahbek, 2011; Matteo, Felicísimo, Pottier, Guisan, & Muñoz, 2012), and methodological issues, such as biases in thresholding (Calabrese, Certain, Kraan, & Dormann, 2013).

Stacked SDMs have been used to study macroecological patterns, including the investigation of abundance patterns (Gomes et al., 2018), biogeographical regions (Amaral, Munhoz, Walter, Aguirre-Gutiérrez, & Raes, 2017; Hazzi, Moreno, Ortiz-Movliav, & Palacio, 2018; Zhang, Slik, & Ma, 2016), variation in latitudinal range size (Garcia-Rosello et al., 2014) and distributional shifts under climate change scenarios (e.g., VanDerWal et al., 2013; Warren et al., 2013). Only a few studies have tested the accuracy of s-SDMs in reproducing these patterns against independent data. The majority of comparisons have investigated the relative prevalence of commission and omission errors at the individual pixel level (e.g., Cooper & Soberón, 2017; Feria & Peterson, 2002; Pineda & Lobo, 2009). Pottier et al. (2013) and D'Amen, Rahbek, Zimmerman, and Guisan (2017) expanded upon this to investigate the distribution of commission and omission errors across geographical and environmental space. However, it is unclear whether omission and commission errors have an effect on downstream macroecological analyses or whether the inclusion of hundreds of SDMs provides a strong enough signal to override errors in some individual models.

More conventional species composition data, for example, in the form of plant species inventories, can be used to delimit biomes, at least in contiguous geographical areas without historical biogeographical barriers that would cause divergent species composition (Mucina, 2019). An advantage of these data relative to s-SDM data is that they are not the output from a model but represent ground truth data. It is possible to fit models to biome classifications obtained through inventory data using environmental data and project these models into areas lacking inventory data (e.g., through the randomForest approach; Breiman, 2001), but this requires a training dataset. This approach can also be used to assess whether environmental data alone can be used to delimit biomes. Inventory data are limited to areas where surveys have taken place and, typically, to woody taxa. Although woody taxa dominate biomass and can be the main component of biome structure, they represent <50% of species richness in all biomes, and their distribution patterns are not representative of non-woody taxa (Droissart et al., 2018).

Here, we test the utility of s-SDMs trained with herbarium specimen data for reproducing macroecological patterns relying upon predictions of species' distributions, focusing on north-east Brazil (NE Brazil). This represents an ideal test case because of the high heterogeneity of the area and the spatial interdigitation of biomes (Silva de Miranda et al., 2018). If the approach can work here, we predict that it will work even better in less biologically complex regions. We test s-SDM performance through comparison with analyses using 1,506 inventories of woody plants. We ask:

1. Can s-SDMs be used to delimit biomes?
2. Are these biomes the same as those identified using inventory data?
3. Are biomes delimited by s-SDMs more accurate than those delimited by environmental data alone?
4. Are biomes delimited by s-SDMs floristically and functionally similar to biomes delimited by inventory data?

Finally, we present a biome map for NE Brazil representing the distributions of 6,134 angiosperm species of diverse growth form to show the potential of s-SDMs in biome delimitation.

2 | METHODS

A full methodological pipeline, showing all input data, analytical steps and results, is shown in Figure 1.

2.1 | Study area

The study area is the political NE of Brazil and Minas Gerais (>2.1 million km^2, referred to here as NE Brazil; Figure 2) and

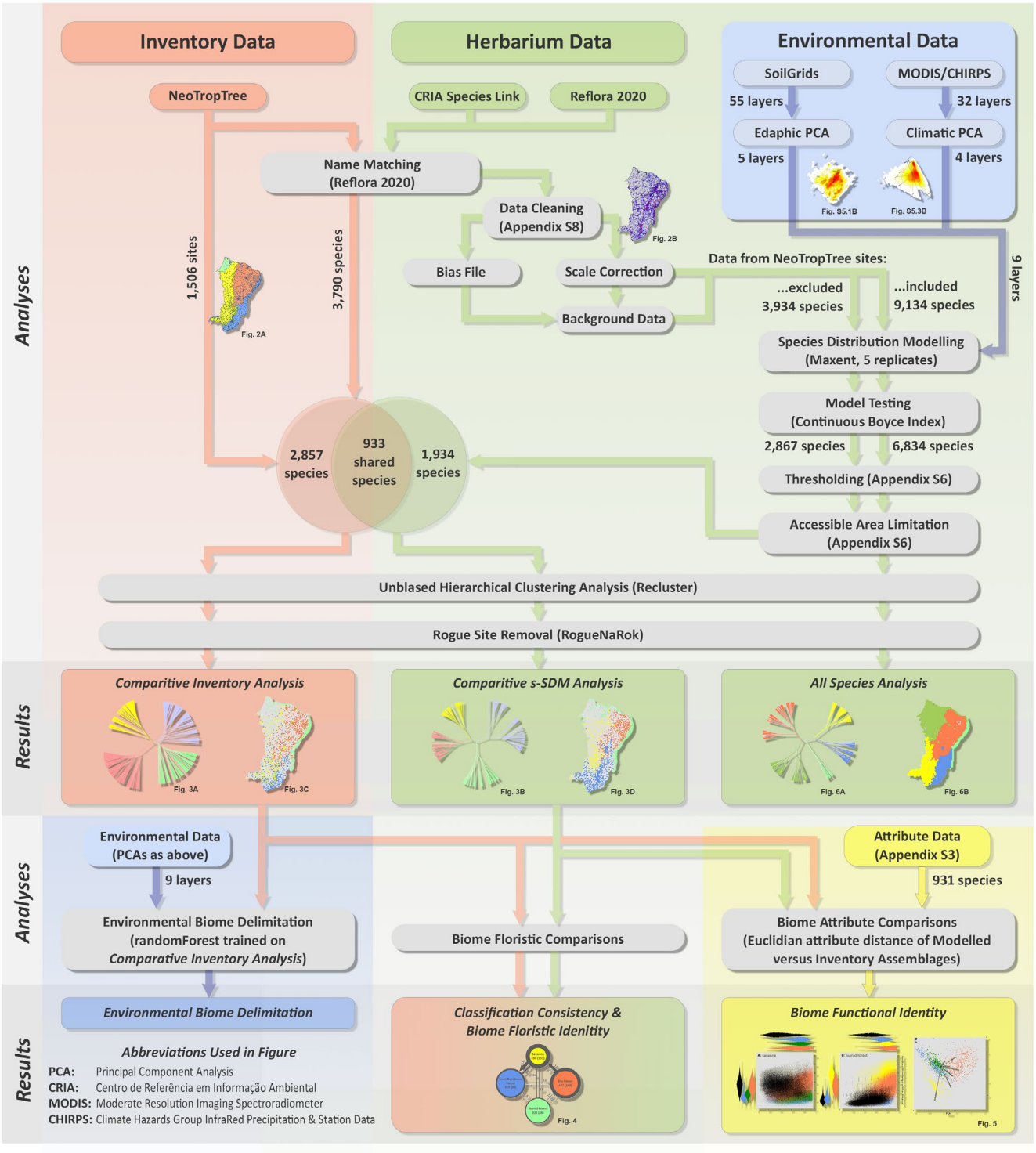

FIGURE 1 Analytical pipeline used in the study. Input data are shown in coloured boxes: red = inventory data; green = herbarium data; blue = environmental data; and yellow = attribute data. Analytical steps are shown in grey boxes. Arrows between boxes indicate the passage of data and are coloured according to the primary source of that data. Results are shown in coloured boxes with italics (headings correspond to those used in the main text). The number of species, sites or layers used in analyses is indicated by arrows where appropriate. Figures 2–6 are provided to aid orientation through the manuscript

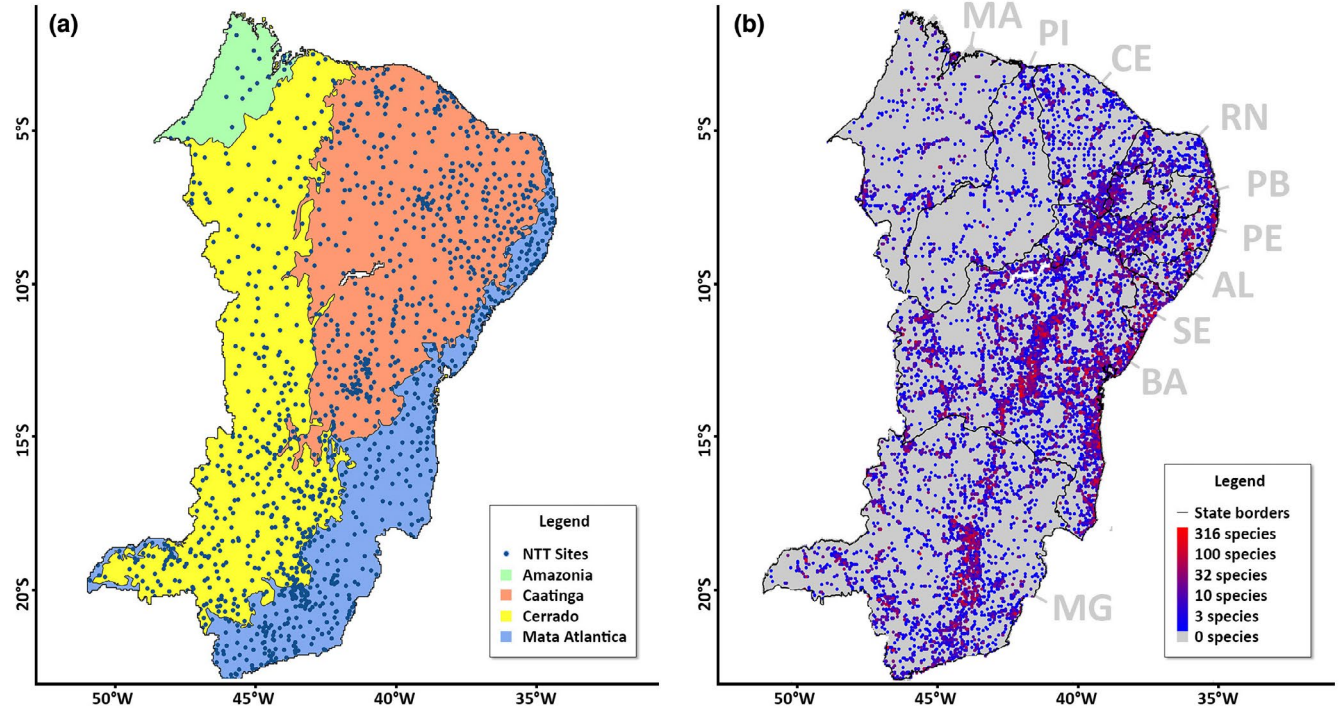

FIGURE 2 The study area of north-east Brazil, including the state of Minas Gerais. (a) NeoTropTree (NTT) sites, showing phytogeographical domains as defined by IBGE (2012). (b) Number of species recorded from herbarium data per 0.05° grid cell after data cleaning ($\log_{10}$ scale), showing state boundaries on the background (AL = Alagoas; BA = Bahia; CE = Ceará; MA = Maranhão; MG = Minas Gerais; PB = Paraíba; PE = Pernambuco; PI = Piauí; RN = Rio Grande do Norte; SE = Sergipe)

was chosen because it encompasses a minimum of three or four major biomes that occur across large climatic gradients. The lowland biomes in NE Brazil include dry forest (or seasonally dry tropical forest; SDTF), savanna, semi-deciduous forest and humid forest (Queiroz, Cardoso, Fernandes, & Moro, 2017; Silva de Miranda et al., 2018). Semi-deciduous forest is intermediate in physiognomy between dry and evergreen humid forest (Dexter et al., 2018), while being floristically closer to humid forest (Bueno et al., 2018; Silva de Miranda et al., 2018). We retain it as a distinct unit, but it could be combined with either dry forest (*sensu* Pennington, Prado, & Pendry, 2000) or humid forest (*sensu* Oliveira-Filho & Fontes, 2000) should greater biome simplicity be desired. Unlike the geographically simple representation of these biomes in the current biogeopolitical map by the Brazilian Institute of Geography and Statistics (IBGE, 2012; Figure 2a), these biomes occur in a complex mosaic across the study region, and their distributions are not yet fully understood at fine spatial scales (Queiroz et al., 2017). This is largely because distinguishing between dry biomes (SDTF and savanna) in remotely sensed land-use maps is challenging because of their superficially similar physiognomy as seen by satellite (mapbiomas.org; Beuchle et al., 2015) and because non-climatic factors, such as fire and small-scale edaphic variation, may control their distribution (Dexter et al., 2018).

2.2 | Data collection

2.2.1 | Inventory data

Inventory data were obtained from the NeoTropTree (NTT) database (Oliveira-Filho, 2017), which includes 1,506 sites in NE Brazil (Figure 2a). NTT data contain presence-only records for trees (defined as freestanding woody plants potentially reaching > 3 m in height), typically based on published ecological inventory or floristic survey data extracted from the literature. NTT data are not systematically collected and do not represent exhaustive surveys, and some sites include as few as eight species. Species records at each NTT site have been supplemented with herbarium records within a 5 km radius of the original inventories (Oliveira-Filho, 2017), and it is not possible to separate these methods. NTT data represent inventories of a single type of vegetation per site, meaning that species and herbarium records found in patches other than the main physiognomy within each site have been excluded (e.g., forest patches within savanna sites) and were used to create a new site that overlaps in geographical space. The latest version of Reflora 2020 (http://floradobrasil.jbrj.gov.br/reflora/listaBrasil/) was used to harmonize the taxonomy with the herbarium data (see Section 2.2.2). About 173 species were excluded from the analysis, most because they are not recorded in the study area according to Reflora 2020 or are not

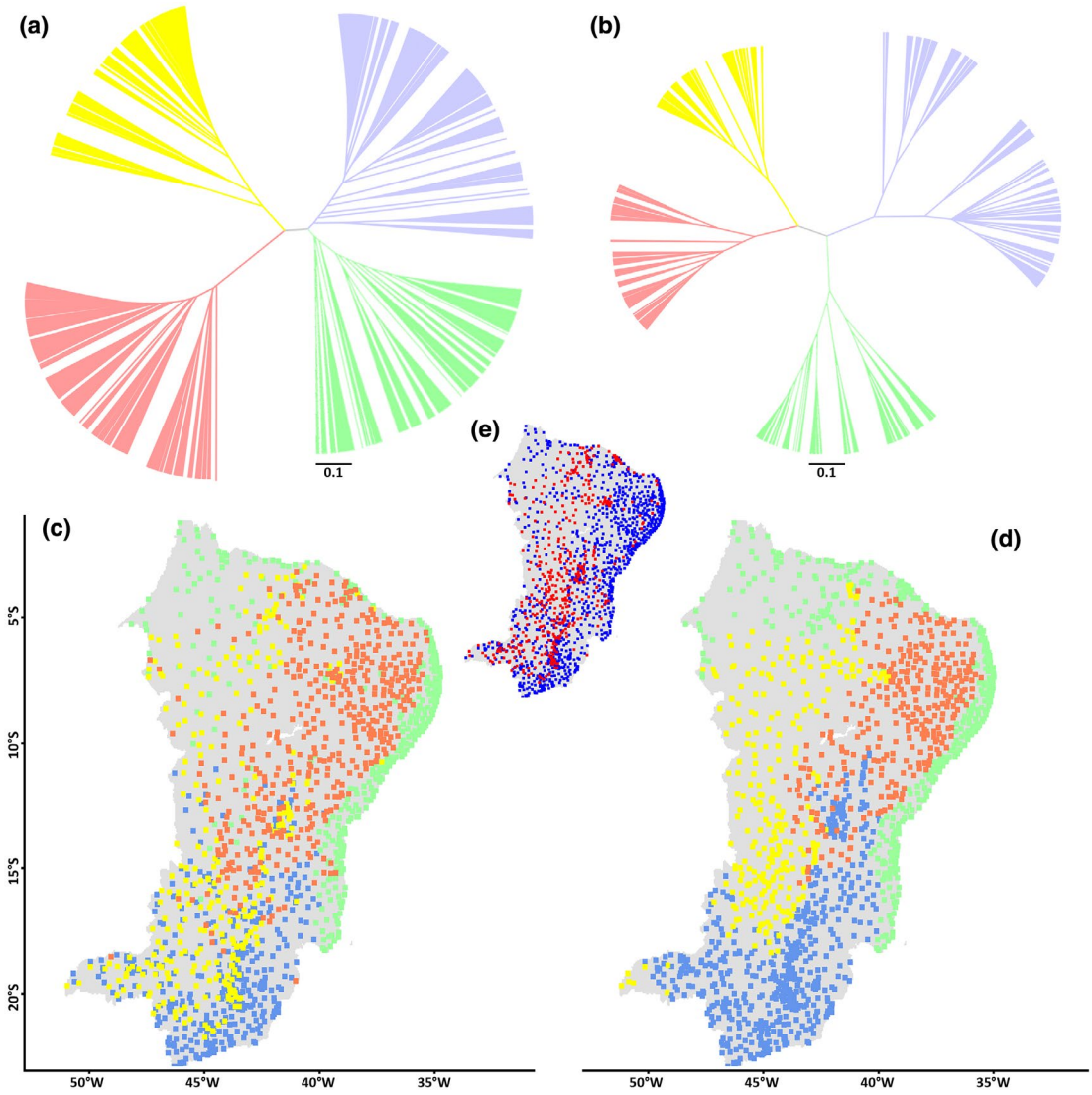

FIGURE 3 Biome classification from inventory data and stacked species distribution models (s-SDMs) across 1,506 NeoTropTree (NTT) sites. (a) Major clusters from the cluster analysis based upon inventory checklists. (b) Major clusters from the cluster analysis based upon s-SDM analysis. (c) Map of major clusters based on checklists. (d) Map of major clusters based on s-SDM analysis. (e) Comparison of checklist and s-SDM results, showing sites that are classified under the same and different biomes in the two analyses in blue and red, respectively. Colours in panels a–d are as follows: green = humid forest; red = dry forest; yellow = savanna; and blue = semi-deciduous forest

angiosperms (Supporting Information Appendix S1). The final inventory dataset included 3,790 angiosperm tree species.

The NTT database contains metadata on vegetation physiognomy and other attributes (e.g., phenology) that can aid in classification of sites to biome. For example, it distinguishes savanna from forest sites, whereas dry forests, semi-deciduous forests and humid forests can be distinguished based on their level of deciduousness. However, these metadata are gleaned from the original sources for the inventory data, and these original sources can be inconsistent in their application of vegetation attribute terminology. Meanwhile, the tree species composition data themselves can be used to classify sites into biomes, which allows for a repeatable, data-driven means of classifying all sites (Silva de Miranda et al., 2018). In practice, these floristically delimited biomes show a strong correspondence to site metadata related to the form and function of vegetation (Dexter et al., 2018). For these reasons, we use floristically delimited biomes for the NTT sites as our baseline for the comparative analyses (see below, Section 2.4).

2.2.2 | Species distribution data

Herbarium data were used as the input for our SDMs. All Brazilian flowering plant occurrence data with original coordinates (i.e., no municipality centroids) and no suspected coordinate issues were downloaded from CRIA Species Link (http://splink.cria.org.br/, October 2017) and the Reflora specimen database (http://floradobrasil.jbrj.gov.br/reflora/herbarioVirtual/, October 2017). The latest version

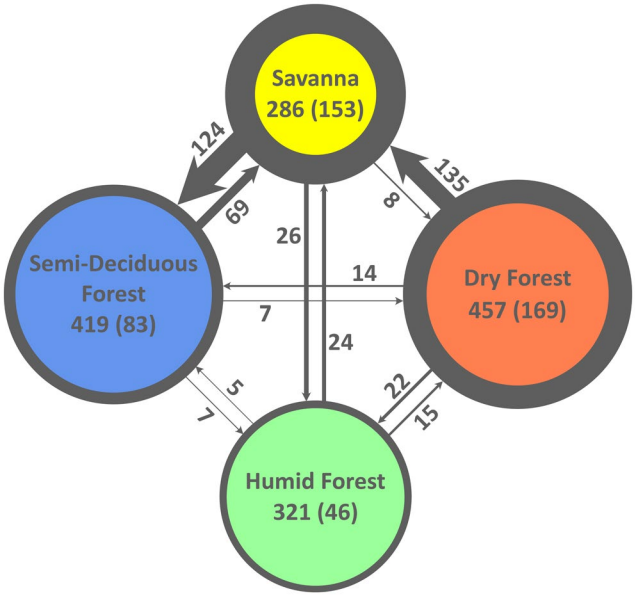

FIGURE 4 Consistency of hierarchical clustering analysis results from inventory and stacked species distribution models (s-SDMs) across 1,483 NeoTropTree (NTT) sites. Four major clusters were recovered in both analyses, corresponding to humid forest (green), dry forest (red), savanna (yellow) and semi-deciduous forest (blue). The size of the grey circle around each biome corresponds to the proportion of sites miscategorized in the s-SDM analysis per biome. The total number of sites is shown within each biome circle, with the number of miscategorized sites in the s-SDM analysis shown in parentheses. Arrows indicate where these sites were miscategorized in the s-SDM analysis. For example, of the 169 miscategorized dry forest sites, the majority (135) were mislabelled as savanna in the s-SDM analysis

of Reflora 2020 was used to harmonize the taxonomy with the inventory data. Data were cleaned in six stages designed to remove records with georeferencing or identification errors (Supporting Information Appendix S2). Environmental bias in occurrence data was addressed through scale correction (also known as spatial filtering) by retaining only a single occurrence record within a 10 km radius for each species, following Kramer-Schadt et al. (2013). Note that although some species have populations outside of Brazil, we chose not to include these records because (a) species distribution data of the quality and quantity available for Brazil typically do not exist for other Neotropical countries, and (b) there are substantial difficulties in matching taxonomic backbones across these datasets. Species with fewer than five records were excluded, and the dataset included 296,439 unique records for 9,134 species.

2.3 | Species distribution modelling

Two sets of SDMs were run using the same settings but based upon different species distribution data. All input data for species with five or more records were used in the all-species analysis (296,439 records for 9,134 species). The comparative s-SDM analysis was designed to be independent of the comparative inventory analysis; therefore, we removed all records that might have been included in the NeoTropTree dataset (i.e., those within 5 km of inventory site centroids). Species with fewer than five records were excluded, and the input data for the comparative s-SDM analysis included 83,509 unique records for 3,934 species. In both analyses, data from across the whole of Brazil were used (i.e., including data from outside of the study area) to minimize niche truncation, which is caused by modelling only a subset of a species range (Austin, 2007). The majority (52%) of species in our analyses are endemic to Brazil; therefore, the entirety of their ranges was modelled.

Climatic and edaphic predictors were used at a 0.05° resolution (*c.* 5.5 km^2; Supporting Information Appendix S3). Climate predictors were derived from remotely sensed temperature (MODIS; Wan, 2014; Wan & Dozier, 1996), rainfall (CHIRPS; Funk et al., 2014) and cloud cover (MODCF; Wilson & Jetz, 2016) data calibrated with ground weather station readings. These climatic data layers outperform those extrapolated from weather stations in modelling plant species distributions (Deblauwe et al., 2016). Edaphic variables from the SoilGrids 250 m database (https://soilgrids.org, February 2017) were used to incorporate edaphic factors expected to be important in controlling species distributions in the study area and which have been found to increase SDM performance in analyses of Neotropical lowland taxa (Figueiredo et al., 2017; Moulatlet et al., 2017).

Climatic (35) and edaphic (55) predictor variables were converted separately into principal components analysis (PCA) axes to reduce the number of predictor variables to four and five, respectively, which in each case explained > 80% of the variation (Supporting Information Appendix S4). This process maximized the inclusion of potentially explanatory data while reducing the number of variables, therefore avoiding issues with collinearity (Dormann et al., 2013) and model overfitting (Peterson, Papeș, & Eaton, 2007).

The SDMs were run in MaxEnt v.3.3.3 using the R package "dismo" (Hijmans, Phillips, Leathwick, & Elith, 2017). We used the MaxEnt default settings, with all feature classes allowed, and with 5-fold cross-validation. Background data were sampled across terrestrial NE Brazil plus circles of 250 km around each species' occurrence points in order to create a biologically realistic model extent, to avoid predicting into areas too far beyond known occurrences and to avoid inflation of model performance metrics through the incorporation of a large number of highly unsuitable background points (Anderson & Raza, 2010). Controlling for bias in sampling effort has been shown to increase the accuracy of SDMs (Stolar & Neilsen, 2015); hence, SDMs were trained with 10,000 background points sampled using an Epanechnikov kernel, following Wiegand and Moloney (2013), and calculated from all angiosperm data for Brazil.

Model performance was evaluated using the continuous Boyce index (CBI; Hirzel, Le Lay, Helfer, Randin, & Guisan, 2006), a presence-only evaluation index based upon the Boyce index (Boyce, Vernier, Nielsen, & Schmiegelow, 2002), calculated with code available at https://rdrr.io/github/adamlilith/enmSdm/man/contBoyce.html. Prediction rasters are split into 10 moving window classes, and

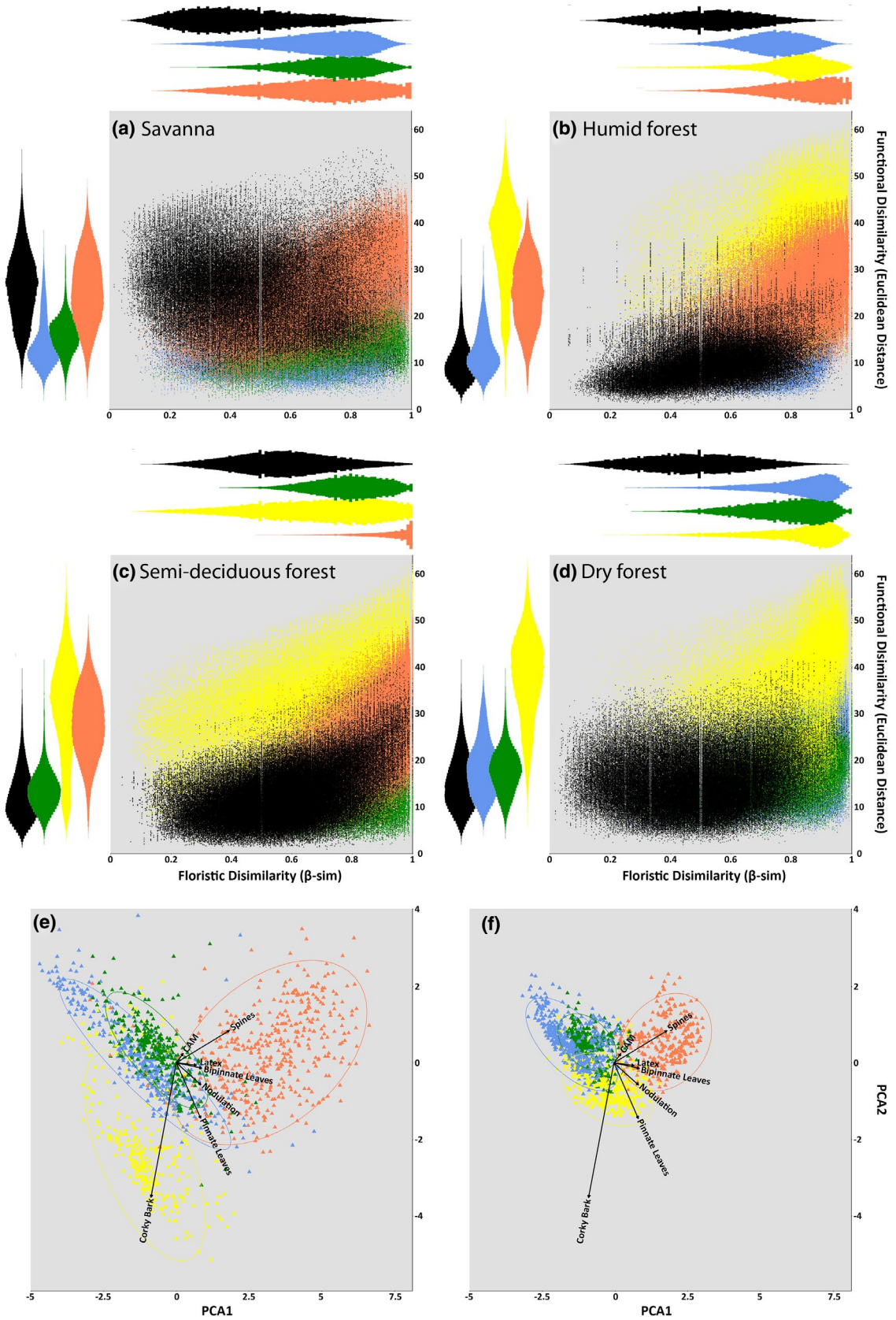

FIGURE 5 (a–d) Floristic and functional similarity of biomes delimited by analyses of stacked species distribution model (s-SDM) data compared with biomes delimited using inventory data. Floristic similarity was measured using Simpson's beta diversity (β-sim) and functional similarity using the Euclidean distance (unitless). (a) Savanna. (b) Humid forest. (c) Semi-deciduous forest. (d) Dry forest. (e,f) Position of assemblages in functional principal components analysis (PCA) space. (e) Inventory data assemblages. (f) s-SDM assemblages. Colours correspond to the following biomes: yellow = savanna; green = humid forest; blue = semi-deciduous forest; and red = dry forests. In each panel, black represents the biome being compared

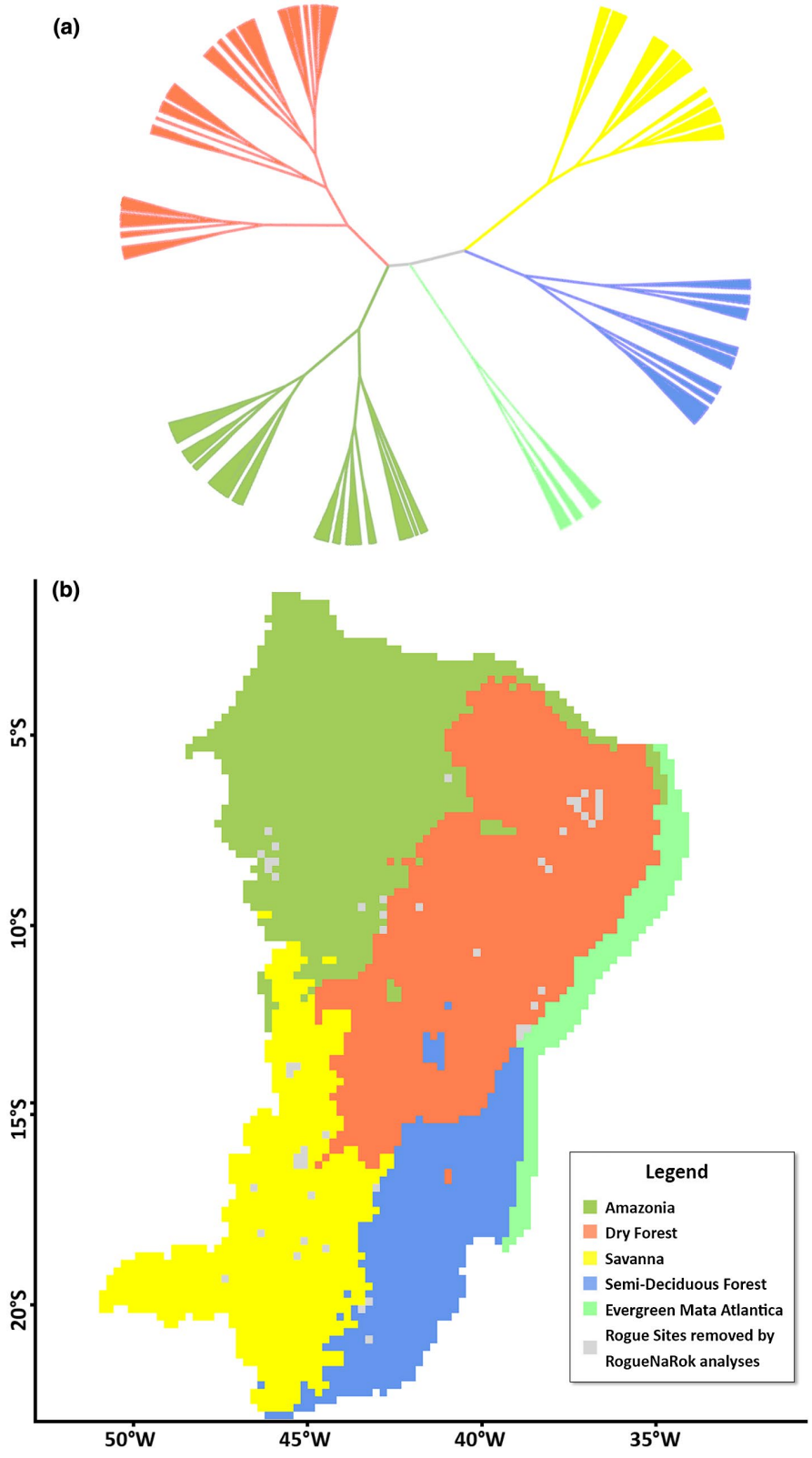

FIGURE 6 Biome classification across north-east Brazil based upon hierarchical clustering analyses of 6,134 species (including trees and other life-forms) at all raster cells in the study area. (a) Major clusters. (b) Map of major clusters

the CBI is the correlation of the ratio of predicted to expected presence evaluation points in each class. The CBI is resistant to stochastic variation at low numbers of presence points (Hirzel et al., 2006). The CBI was calculated for each of the five model replicates, and species with a mean CBI < 0.25 were discounted from further analyses. This is a relatively conservative criterion, because models with a CBI > 0 perform better than random.

Robust SDMs with CBIs ≥ 0.25 were produced for 6,834 species for the models produced for the all-species analysis (75%). A slightly lower percentage of the SDMs produced for the comparative s-SDM

analysis were deemed acceptable, totalling 2,867 species (73%). SDMs were converted from probability distributions to binary maps with a 10 percentile training presence logistic threshold, which was chosen because it is the most strict of the commonly used threshold values. A posteriori accessible area limitation was used to exclude areas likely to represent commission areas (Supporting Information Appendix S5). Predicted species lists were estimated for every raster cell in NE Brazil for the all-species analysis. For the comparative s-SDM analysis, species lists were estimated for the 1,506 NTT sites.

2.4 | Biome delimitation

Our approach relies upon the assumption that floristic patterns are a good proxy for functional patterns, because we lack sufficient functional data at the subcontinental scale for use in biome delimitation. Three biome delimitation analyses were run using the same methods but based upon different species lists. For the comparative analyses, we used lists of the same 933 species at the 1,506 inventory sites. The 933 species were those for which we could produce acceptable SDMs and were found in the inventory data. Lists used in the comparative inventory analyses were based upon the presence of these species in the inventory data; lists used in the comparative s-SDM analysis were based upon their predicted presence in thresholded s-SDMs. For the all-species analysis, we used predicted presence–absence lists for 6,834 species at each raster cell in NE Brazil estimated from thresholded s-SDMs. Owing to computational limitations, raster cells were aggregated to a resolution of 0.25°, and species were scored as present or absent in each cell based upon their predicted presence in any of the subcells at the original 0.05° resolution.

In each biome delimitation analysis, hierarchical cluster analysis based on β-diversity (Simpson's dissimilarity) was performed based upon the species lists. Unbiased cluster analysis was used, where each analysis was subjected to randomizing of the row order in matrices 100 times using the "recluster" package in R (Dapporto et al., 2013, 2015). RogueNaRok was used to remove rogue sites responsible for reducing resolution in the resulting 50% majority rule consensus dendrogram (19 sites in comparative inventory analysis, seven in comparative s-SDM analysis and 61 cells in all-species analysis; Supporting Information Appendix S6; Aberer, Krompass, & Stamatakis, 2012).

Resultant clusters were split into biomes based upon a process of reciprocal illumination, by considering the overall shape (branching patterns) of the trees, our biological knowledge of vegetation patterns in the study area and consultation with additional experts (M. F. Fernandes, M. Moro, D. Neves, F. Pezzini, L. P. Quieroz, R. M. Santos and DryFlor, pers. comm.). Other studies have taken a similar approach (Fayolle et al., ; Silva de Miranda et al., 2018). Alternative approaches, such as k-means methodologies (e.g., Amaral et al., 2017), also rely upon the knowledge of specialists to define an a priori number of biomes, and analytical workflows based on fully objective criteria still rely upon the authors checking their obtained biomes against specialists' knowledge (e.g., Edler, Guedes, Zizha, Rosvall, & Antonelli, 2017). Our priorities were the generality of our results and maximizing the similarity of our mapped clusters to the IBGE (2012) classification, which recognizes four biomes in NE Brazil. We therefore focused on delimiting biomes in the broadest sense. For all groups, we considered further bifurcations, but these splits were rejected based upon expert knowledge.

2.5 | Biome naming and floristic identity

The names of recovered biomes were determined by their mapped distributions with reference to expert knowledge (ourselves and additional experts; see previous subsection), the IBGE (2012) classification and the NTT site metadata (Oliveira-Filho, 2017). To test whether inventory data and s-SDMs delimit similar biomes, we compared the classification of individual sites between analyses of inventory and s-SDM data.

To determine whether biomes delimited by s-SDMs were similar to their counterparts delimited by inventory data in terms of floristic composition, we used an approach based upon Simpson's β-diversity (β-sim) implemented in the R package "betapart" (Baselga, 2010; Baselga, Orme, Villéger, De Bortoli, & Leprieur, 2018). β-sim was used because it measures only floristic turnover between sites and does not incorporate nestedness. Nestedness was discounted because s-SDMs predicted more species per site (mean = 217.8) than found in inventory surveys (mean = 95.9), which represent non-exhaustive surveys with variable sampling effort (Oliveira-Filho, 2017). We calculated β-sim from each s-SDM assemblage to every inventory data assemblage. We then carried out a Kruskal–Wallis test to determine whether within-biome β-sim values were smaller than across-biome comparisons.

2.6 | Environmental biome delimitation

To assess whether s-SDM data were able to identify biomes better than environmental data alone, we used a random forest classification tree approach (Breiman, 2001) in the "randomForest" package in R (Liaw & Wiener, 2002). We used the same PCAs of environmental variables that we used as predictor variables for SDMs (Supporting Information Appendix S3). We carried out two random forest analyses: one based upon the edaphic and climatic PCA layers and one based upon the climatic PCA layers alone. The results of these analyses include an estimated error rate in prediction of the biome identity of sites given the imputed environmental data. In order to determine the success of the random forest classification tree approach, we compared the error rates with those from the comparative inventory analysis.

2.7 | Biome functional identity

Presence or absence of six independent plant attributes was recorded for 931 species (latex, corky bark, spines, compound leaves,

nodulation and Crassulacean acid metabolis photosynthesis; Supporting Information Appendix S7). Attributes were chosen based upon (a) their hypothesized strong links to environmental and functional differences among biomes in NE Brazil, and (b) ease of recording from herbarium specimens, floristic treatments and taxonomic monographs (Supporting Information Appendix S7).

To examine the functional identity and similarity of biomes, we first calculated the proportion of species at each site with each attribute. To determine whether biomes delimited by the comparative s-SDM analysis were different from each other in terms of attributes, we performed a PCA upon these data. We repeated this with the assemblages in the inventory dataset. To determine whether the comparative s-SDM analysis biomes were functionally similar to corresponding biomes delimited by the comparative inventory analysis, we grouped s-SDM assemblages by biomes. We then calculated the functional Euclidean distance from every s-SDM assemblage to every inventory data assemblage. We carried out a Kruskal–Wallis test to determine whether within-biome Euclidean distances were smaller than across-biome comparisons. The purpose of this test was to confirm whether the delimited biomes differ in functional attribute space.

3 | RESULTS

3.1 | Species distribution modelling

The mean CBI value for the 933 SDMs used in the comparative s-SDM analysis was 0.51, indicating high model performance. The area under the curve and CBI were positively correlated with each other and with the number of specimens (Supporting Information Appendix S8). All model statistics and the contribution of each environmental PCA axis to the models are provided in the Supporting Information (Supplementary Information Appendix S8).

The mean number of species per site in inventories was 96 and the mean predicted per site by s-SDMs was 218. The mean number of species predicted present by s-SDMs but absent in inventory data was 160. The mean number of species predicted absent by s-SDMs but present in inventory data was 38 (omission errors). In total, s-SDMs correctly predicted the presence of 61% of species in the inventory data.

3.2 | Comparative analyses

3.2.1 | Biome delimitation

Hierarchical clustering analysis of the inventory data (the comparative inventory analysis) produced four higher level groups (Figure 3a). We identified an Amazon biome, here labelled as humid forest, encompassing 321 sites in northern Maranhão and Piauí states and the wetter elements of the Atlantic forest close to the eastern coast of NE Brazil. This grouping also includes some sites in southern Maranhão, Bahia and Paraíba (Figure 3c). The three other biomes are comparatively dry: dry forest, semi-deciduous forest and savanna (Figure 3c). These three biomes are distributed throughout the interior of north-east Brazil, largely overlapping with some of the extent of the Caatinga, Mata Atlantica and Cerrado biogeopolitical domains of IBGE (2012), respectively (Figure 2a). The error rate in assigning sites to these biomes was determined as 27% using climatic data and 29% using climatic data in combination with edaphic data through randomForest analysis.

The comparative s-SDM analysis also identified four biomes (Figure 3b), which correspond to the same four biomes identified by inventory data based upon their geographical distributions (Figure 3d). However, the s-SDM results show more geographically aggregated biomes compared with the inventory data, which show more spatial interdigitation, particularly within the western part of the study area (Figure 3c).

The majority of sites were assigned to the same biome in both analyses (1,032 sites, 70%; Supporting Information Appendix S9), and levels of misclassification among three of the four biome types were low (Figure 4). We found very little error in classification between the humid forest biome and other biomes. Furthermore, s-SDMs were readily able to distinguish the semi-deciduous forest and dry forest biomes. The majority of site classification errors (335 sites, 75% of total errors) were between the savanna and the semi-deciduous or dry forest biomes. In other words, s-SDM analyses can distinguish between wet and seasonally dry biomes and between dry and semi-deciduous forest, but not between the savanna and the two seasonally dry forest biomes.

3.2.2 | Biome floristic and functional identity

All four biomes as delimited by s-SDMs were significantly more floristically similar to their corresponding biome in analyses of inventory data than to other biomes (Figure 5a–d). This was confirmed with a Kruskal-Wallis test for each biome, which in each case was highly significant ($p < .0001$).

The four biomes as delimited by inventory data showed varying degrees of functional differentiation in PCA space (Figure 5e). Semi-deciduous forest and humid forest had the most overlap, whereas dry forest and savanna each displayed a unique and broad distribution in PCA space. This pattern was replicated by the s-SDM biomes (Figure 5f), but each biome had a narrower distribution in PCA space.

The biomes as delimited by inventory data were, for the most part, well differentiated in PCA space (Figure 5e). Savanna was differentiated primarily by corky bark, whereas dry forest was differentiated by a combination of spines and bipinnate leaves. The dry and humid forest biomes showed less differentiation and were categorized primarily by the low number of species with any of the six attributes examined. The biomes as delimited by s-SDMs trended in the same way in PCA space as the inventory data biomes, but in general were less functionally diverse (Figure 5f).

Three of the four biomes as delimited by s-SDMs were significantly more functionally similar to their corresponding biome than other biomes in analyses of inventory data (Figure 5a–d). These three biomes were the dry forest, semi-deciduous forest and the humid forest. However, the modelled savanna biome was significantly less similar to the savanna in analyses of inventory data than all other biomes (Figure 5a).

3.3 | All-species analyses

The mean CBI for the 6,823 SDMs included in the analysis was 0.54. Hierarchical clustering analysis of species lists for all raster cells for these analyses produced five higher level biome groups (Figure 6), after the removal of 61 sites through RogueNaRok analyses. These groups were largely concordant with those recognized above (Figure 3), with two principal differences. First, the evergreen coastal part of the humid forest biome, here referred to as Mata Atlantica, was recognized as its own group, separate from humid forest in the northwest of the study area (i.e., Amazonia; Figure 6). Second, savanna reached less far north, where it is replaced by the remainder of the humid forest biome (Amazonia; Figure 6).

4 | DISCUSSION

This study explored the power of s-SDMs relative to inventory data in recovering macroecological patterns and is the first attempt to test the utility of s-SDMs in delimiting biomes. Many previous studies have tested the performance of s-SDMs in predicting species richness per site (e.g., D'Amen et al., 2015, 2017; Feria & Peterson, 2002), but here we focused on testing the implication of these differences for downstream analyses of floristic composition and functional diversity patterns at large spatial scales.

4.1 | Biome delimitation and identity

The concept of biomes has varied through time, but all definitions have shared the aim to define broad-scale, ecologically meaningful units of vegetation. The approach taken to delimit these units has varied greatly, however, with emphasis on the floristic, structural or functional components of vegetation (Mucina, 2019). Given that global, continental or even regional scale datasets on these factors have been and remain rare, most existing biome maps have been based on expert views (e.g., IBGE, 2012) or the biome distributions have been estimated by their hypothesized environmental determinants (e.g., Whittaker, 1970).

Stacked SDMs provide a potential solution for biome mapping by providing continuous floristic data at large spatial scales that can be linked to species functional traits and attributes. s-SDMs allow us to generate the spatially continuous estimates of species distributions from species distribution data. Spatially continuous distribution data are required for the production of data-driven, repeatable maps of biome distributions. Given that s-SDMs are known to both over- and under-predict species distributions, it has remained unclear whether data from s-SDMs are of sufficient quality to delimit ecologically meaningful biomes.

Here, we compared biomes delimited with s-SDM and inventory data to test whether s-SDM can reproduce macroecological patterns at large spatial scales. Identical hierarchical clustering analyses were run on s-SDM and inventory data for the same 933 species and 1,506 sites; therefore, all the difference between the results of these analyses must result from differences in the species lists derived from s-SDMs and inventory data. The species lists provided by the s-SDMs contain both commission and omission errors. There are a number of potential sources for these errors in both the NTT dataset and the s-SDM dataset. Those in the NTT data include misidentifications, the often-incomplete nature of NTT surveys and the exclusion of species found in patches other than the main physiognomy within each NTT site (see Section 2). There are numerous causes of omission and commission errors (as described in the Section 1 and citations therein). Omission and commission errors are inevitable in any s-SDM study; therefore, their inclusion in the present study aids in testing the efficacy of s-SDMs in biome delimitation in a real-world setting.

It is possible that our choice of SDM input data and the methodological choices we made in producing our s-SDMs have had a significant impact upon the biomes identified by both the comparative s-SDM analysis and the all-species analysis. Inherent in any SDM study are numerous methodological choices, including but not limited to: choice of input, environmental and background data; data-cleaning pipelines; methods for minimizing bias in input data (Kramer-Schadt et al., 2013; Wiegand & Moloney, 2013); model algorithm and parameters (Hijmans et al., 2017); model extent (Anderson & Raza, 2010); model testing (Hirzel et al., 2006); and threshold values (Calabrese et al., 2013). No standard SDM pipeline exists, and it is beyond the scope of this paper to test the effect of our choices on our results. We have instead opted for a single pipeline that approximates "best practice" given the limitations of our data and study area. This gives the comparative s-SDM analysis the best possible chance of reproducing the results of the comparative inventory analysis.

The comparative s-SDM analysis and comparative inventory analysis both resulted in the recognition of four biomes similar in geographical distributions and floristic composition to those detected in previously published analyses (Queiroz et al., 2017; Silva de Miranda et al., 2018) and, in most cases, their functional identity. This indicates that s-SDMs are able to recover macroecological patterns with accuracy. The majority of sites were assigned to the same biome in both analyses (1,032 sites, 70%), and levels of classification errors among the majority of biome types were low. We found little error in classification between the humid forest biome and other biomes. Furthermore, s-SDMs were readily able to distinguish semi-deciduous forest and dry forest biomes. The majority of classification errors (335 sites, 75% of the total)

were between the savanna biome and the two other seasonally dry biomes (semi-deciduous and dry forest). In other words, s-SDM analyses can predict the boundaries of humid and seasonally dry biomes but not the boundaries of the savanna versus seasonally dry forested biomes.

The overall site classification error rate in the comparative s-SDM analysis (30%) is similar to that when classifying sites based on climatic data alone (27%) or on climatic and edaphic data in combination (29%) in the randomForest analyses. We note, however, that other methods of classifying biomes using environmental data might perform differently. A significant disadvantage of randomForest analyses is that they require a training dataset (in this case, site classifications determined through analysis of inventory data) and are therefore not applicable to areas lacking inventory data.

4.2 | Modelling savannas

Our results show that analyses of s-SDM data struggle to identify the distribution and floristic and functional composition of savanna sites (Figures 2–4). Our SDMs are primarily based upon climatic data, and this therefore implies that savanna overlaps with both dry and semi-deciduous forest in climatic space; thus, climate cannot be used to predict the distributions of savanna. There is considerable debate over the primary factors governing the distributions of savannas, but climate is a poor predictor of South American savanna distributions (Lehmann et al., 2014). Rather, debate focuses on whether fire, disturbance or soils are the primary factors governing the distributions of savannas (e.g., Hoffmann et al., 2012; Veenendaal et al., 2018), none of which can be captured well in SDMs in the Tropics at present. Although we included soil layers coming from SoilGrids (Hengl et al., 2014) in our SDMs and these add statistical power to our models (Supporting Information Appendix 4. Table S4.1), such layers do not fully capture the edaphic complexity of our study area, unfortunately. No suitable fire layer currently exists for use in SDMs, although remotely sensed fire data might eventually fulfil this purpose. Our s-SDMs lack the predictive variables required to capture the full spectrum and complexity of variables implicated in the assemblage of savanna communities.

Although s-SDMs are weak at distinguishing the distribution of the savanna biome, the modelled savanna biome is significantly more floristically similar to the savanna biome derived from inventory data, but functionally more similar to both the humid and dry forest biome than the savanna biome derived from the inventory data (Figure 5a). The savanna biome differs functionally from the other three biomes primarily in its high proportion of species with corky bark (Figure 5e), but modelled savanna assemblages have much lower percentages of corky-barked species (Figure 5f). The functional differences between s-SDMs and inventory data in the fire-prone savanna assemblages are likely to be attributable to environmental filtering for corky-barked species, which is not encapsulated in SDMs lacking a fire layer, indicating that fire plays a strong role in the community assembly of savannas.

A further complication in modelling savanna distribution is that small patches of savanna, semi-deciduous forest and dry forest often occur in complex, interdigitated matrices (DryFlor, 2016; Silva de Miranda et al., 2018), and individual patches are often smaller than the *c*. 5.5 km^2 resolution of our s-SDMs. This reduces the statistical power of s-SDMs to distinguish the environmental requirements of savanna species.

4.3 | All-species analyses

Perhaps the most compelling argument for using s-SDMs in macroecological analyses is their ability to extend our knowledge of biodiversity patterns beyond trees. At present, inventory data are largely limited to trees. Consequently, our understanding of biodiversity patterns in plants is biased towards a single growth form that represents < 50% of all known plant species across all biomes (Cardoso et al., 2017; Droissart et al., 2018). Given that distribution patterns, range sizes and the spatial patterns of functional traits in relationship to climate differ significantly among life-forms (Droissart et al., 2018; Šímová et al., 2018; Xu et al., 2017), modelling of all growth forms with a more representative sample of species in s-SDMs is a significant advantage in analyses aiming to understand overall macroecological patterns.

Our all-species biome map strongly supports this argument. Through the inclusion of non-tree species in s-SDM biome delimitation analyses, we were able to distinguish two floristic units within the humid forest biome: the Evergreen Mata Atlantica and Amazonia (Figure 5a,b). Tree distributions are likely not to be a good proxy for the distributions of all macroecological patterns and processes, and by neglecting other life-forms, our views of macroecological patterns might be significantly biased. Floristic inventories of non-woody taxa are in their infancy or limited in extent; thus, s-SDMs are currently the only way of investigating macroecological patterns for all life-forms while retaining species identity across large spatial scales. The output of s-SDM analyses, however, need to be treated with caution where biome distributions are not determined by climate alone.

4.4 | Conclusions

Stacked SDMs are an increasingly important tool in the investigation of macroecological patterns, but their efficacy remains largely untested. We investigated the utility of s-SDMs in predicting geographical, floristic and functional characteristics of biomes at 1,506 sites across NE Brazil compared with an inventory dataset. The s-SDMs recovered the same broad-scale biomes as inventory data, with similar geographical, floristic and functional characteristics, but they struggled in areas where non-climatic factors, such as fire or soil, play key roles in governing distributions. This is likely to be because s-SDMs do not include the most important variables governing community assembly processes for savanna systems, and we

thus recommend caution in macroecological analyses using s-SDMs in these areas. s-SDMs can, however, be used in areas with few or no floristic inventories and to study the distributions of taxa not usually included in such studies. Given these significant advantages and their generally low error rates, our study demonstrates that s-SDMs can be used to elucidate functional and floristic macroecological patterns with confidence even in the majority of complex, tropical settings.

ACKNOWLEDGEMNTS

This paper, P.W.M., T.E.S., D.C. and R.T.P. were funded by the Natural Environment Research Council-Newton grant NE/N012526/1 "Nordeste: New Science for a Neglected Biome". D.C. and G.R. were funded by the Royal Society Advanced Fellowship grant NAF/R1/180331, Fundação de Amparo à Pesquisa da Bahia (Universal no. APP0037/2016), and Conselho Nacional de Desenvolvimento Científico e Tecnológico Research Productivity PQ-2 grant 308244/2018-4. P.L.M.S. was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior grant 99999.013197/2013-04 under the Science without Borders Programme and the Université de Liège post-doctoral fellowship under the IPD-STEMA scheme (2019). We thank Centro de Referência em Informação Ambiental (CRIA) and Reflora for sharing their complete distribution datasets.

DATA AVAILABILITY STATEMENT

The inventory data used to produce this paper can be accessed freely at http://www.neotroptree.info/, and the species occurrence data can be accessed at http://splink.cria.org.br/ and http://floradobrasil.jbrj.gov.br/herbariovirtual/

ORCID

Peter W. Moonlight https://orcid.org/0000-0003-4342-2089
Pedro Luiz Silva de Miranda https://orcid.org/0000-0002-3008-1556
Domingos Cardoso https://orcid.org/0000-0001-7072-2656
Kyle G. Dexter https://orcid.org/0000-0001-9232-5221
R. Toby Pennington https://orcid.org/0000-0002-8196-288X
Gustavo Ramos https://orcid.org/0000-0002-4539-394X
Tiina E. Särkinen https://orcid.org/0000-0002-6956-3093

REFERENCES

Aberer, A. J., Krompass, D., & Stamatakis, A. (2012). Pruning rogue taxa improves phylogenetic accuracy. *Systematic Biology*, *62*, 162–166.

Amaral, A. G., Munhoz, C. B. R., Walter, B. M. T., Aguirre-Gutiérrez, J., & Raes, N. (2017). Richness pattern and phytogeography of the Cerrado herb–shrub flora and implications for conservation. *Journal of Vegetation Science*, *28*, 848–858. https://doi.org/10.1111/jvs.12541

Anderson, R. P., & Raza, A. (2010). The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: Preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. *Journal of Biogeography*, *37*, 1378–1393. https://doi.org/10.1111/j.1365-2699.2010.02290.x

Austin, M. P. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, *200*, 1–19. https://doi.org/10.1016/j.ecolmodel.2006.07.005

DryFlor (2016). Plant diversity patterns in neotropical dry forests and their conservation implications. *Science*, *353*, 1383–1387. https://doi.org/10.1126/science.aaf5080

Baselga, A. (2010). Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, *19*, 134–143. https://doi.org/10.1111/j.1466-8238.2009.00490.x

Baselga, A., Orme, D., Villéger, S., De Bortoli, J., & Leprieur, F. (2018). Package betapart: Partitioning beta diversity into turnover and nestedness. R package version 1.5.0. Retrieved from https://CRAN.R-project.org/package=betapart

Beuchle, R., Grecchi, R. C., Shimabukuro, Y. E., Seliger, R., Eva, H. D., Sano, E., & Achard, F. (2015). Land cover changes in the Brazilian Cerrado and Caatinga biomes from 1990 to 2010 based on a systematic remote sensing sampling approach. *Applied Geography*, *58*, 116–127. https://doi.org/10.1016/j.apgeog.2015.01.017

Boyce, M. S., Vernier, P. R., Nielsen, S. E., & Schmiegelow, F. K. A. (2002). Evaluating resource selection functions. *Ecological Modelling*, *157*, 281–300. https://doi.org/10.1016/S0304-3800(02)00200-4

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Bueno, M. L., Dexter, K. G., Pennington, R. T., Pontara, V., Neves, D. M., Ratter, J. A., & Oliveira-Filho, A. T. (2018). The environmental triangle of the Cerrado domain: Ecological factors driving shifts in tree species composition between forests and savannas. *Journal of Ecology*, *106*, 2109–2120. https://doi.org/10.1111/1365-2745.12969

Calabrese, J. M., Certain, G., Kraan, C., & Dormann, C. F. (2013). Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography*, *23*, 99–112. https://doi.org/10.1111/geb.12102

Cardoso, D., Särkinen, T., Alexander, S., Amorim, A. M., Bittrich, V., Celis, M., … Forzza, R. C. (2017). Amazon plant diversity revealed by a taxonomically verified species list. *Proceedings of the National Academy of Sciences USA*, *114*, 10695–10700. https://doi.org/10.1073/pnas.1706756114

Castro-Insua, A., Gomez-Rodrıguez, C., & Baselga, A. (2016). Break the pattern: Breakpoints in beta diversity of vertebrates are general across clades and suggest common historical causes. *Global Ecology and Biogeography*, *25*, 1279–1283. https://doi.org/10.1111/geb.12507

Cooper, J. C., & Soberón, J. (2017). Creating individual accessible area hypotheses improves stacked species distribution model performance. *Global Ecology and Biogeography*, *27*, 156–165. https://doi.org/10.1111/geb.12678

D'Amen, M., Pradervand, J.-N., & Guisan, A. (2015). Predicting richness and composition in mountain insect communities at high resolution: A new test of the SESAM framework. *Global Ecology and Biogeography*, *24*, 1443–1453. https://doi.org/10.1111/geb.12357

D'Amen, M., Rahbek, C., Zimmerman, N. E., & Guisan, A. (2017). Spatial predictions at the community level: From current approaches to future frameworks. *Biological Reviews of the Cambridge Philosophical Society*, *92*, 169–187. https://doi.org/10.1111/brv.12222

Dapporto, L., Ramazzotti, M., Fattorini, S., Talavera, G., Vila, R., & Dennis, R. L. H. (2013). Package recluster: An unbiased clustering procedure for beta-diversity turnover. *Ecography*, *36*, 1070–1075. https://doi.org/10.1111/j.1600-0587.2013.00444.x

Dapporto, L., Ramazzotti, M., Fattorini, S., Talavera, G., Vila, R., & Dennis, R. L. H. (2015). Package recluster: Ordination methods for the analysis of beta-diversity indices. R. package version 2.8. Retrieved from https://CRAN .R-project.org/packages=recluster

Deblauwe, V., Droissart, V., Bose, R., Sonké, B., Blach-Overgaard, A., Svenning, J.-C., … Couvreur, T. L. P. (2016). Remotely sensed

temperature and precipitation data improve species distribution modelling in the tropics. *Global Ecology and Biogeography*, *25*, 443–454. https://doi.org/10.1111/geb.12426

Dexter, K. G., Pennington, R. T., Oliveira-Filho, A. T., Bueno, M. L., Silva de Miranda, P. L., & Neves, D. M. (2018). Inserting tropical dry forests into the discussion on biome transitions in the tropics. *Frontiers in Ecology and Evolution*, *6*, 1–7. https://doi.org/10.3389/fevo.2018.00104

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*, 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Droissart, V., Dauby, G., Hardy, O. J., Deblauwe, V., Harris, D. J., Janssens, S., … Couvreur, T. L. P. (2018). Beyond trees: Biogeographical regionalization of tropical Africa. *Journal of Biogeography*, *45*, 1153–1167. https://doi.org/10.1111/jbi.13190

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J.-P., & Guisan, A. (2011). Predicting spatial patterns of plant species richness: A comparison of direct macroecological and spatial species stacking approaches. *Diversity and Distributions*, *17*, 1122–1131.

Edler, D., Guedes, T., Zizha, A., Rosvall, M., & Antonelli, A. (2017). Infomap bioregions: Interactive mapping of biogeographical regions from species distributions. *Systematic Biology*, *66*, 197–204. https://dx.doi.org/10.1093%2Fsysbio%2Fsyw087

Fayolle, A., Swaine, M. D., Aleman, J., Azihou, A. F., Bauman, D., te Beest, M., … Woollen, E. (2018). A sharp floristic discontinuity revealed by the biogeographic regionalization of African savannas. *Journal of Biogeography*, *46*, 454–465. https://doi.org/10.1111/jbi.13475

Feria, A. P., & Peterson, A. T. (2002). Prediction of bird community composition based on point-occurrence data and inferential algorithms: A valuable tool in biodiversity assessments. *Diversity and Distributions*, *8*, 49–56. https://doi.org/10.1046/j.1472-4642.2002.00127.x

Figueiredo, O. G., Zuquim, G., Tuomisto, H., Moulatlet, G. M., Balslev, H., & Costa, F. R. C. (2017). Beyond climate control on species range: The importance of soil data to predict distribution of Amazonian plant species. *Journal of Biogeography*, *45*, 190–200. https://doi.org/10.1111/jbi.13104

Funk, C. C., Peterson, P. J., Landsfeld, M. F., Pedreros, D. H., Verdin, J. P., Rowland, J. D., … Verdin, A. P. (2014). A quasi-global precipitation time series for drought monitoring. *U.S. Geological Survey Data Series*, *832*, 1–4.

García-Roselló, E., Guisande, C., Manjarrés-Hernández, A., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., … Lobo, J. M. (2014). Can we derive macroecological patterns from primary Global Biodiversity Information Facility data? *Global Ecology and Biogeography*, *24*, 335–347. https://doi.org/10.1111/geb.12260

Gomes, V. H. F., Ijff, S. D., Raes, N., Amaral, I. L., Salomão, R. P., de Souza Coelho, L., ter Steege, H. (2018). Species distribution modelling: Contrasting presence-only models with plot abundance data. *Scientific Reports*, *8*, 1003. https://doi.org/10.1038/s41598-017-18927-1

Gould, W. (2000). Remote sensing of vegetation, plant species richness, and biodiversity hotspots. *Ecological Applications*, *10*, 1861–1870.

Guisan, A., & Rahbek, C. (2011). SESAM – a new framework for predicting spatio-temporal patterns of species assemblages: Integrating macroecological and species distribution models. *Journal of Biogeography*, *38*, 1433–1444.

Hazzi, N. A., Moreno, J. S., Ortiz-Movliav, C., & Palacio, R. D. (2018). Biogeographic regions and events of isolation and diversification of the endemic biota of the tropical Andes. *Proceedings of the National Academy of Sciences USA*, *115*, 7985–7990. https://doi.org/10.1073/pnas.1803908115

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., … Gonzalez, M. R. (2014). SoilGrids1km — Global soil information based on automated mapping. *PLoS One*, *9*, e114788. https://doi.org/10.1371/journal.pone.0105992

Hijmans, R. J., Phillips, S., Leathwick, J., & Elith, J. (2017). Package dismo: species distribution modelling. R package version 1.1-4. Retrieved from https://CRAN.R-project.org/package=dismo

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, *199*, 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017

Hoffmann, W. A., Geiger, E. L., Gotsch, S. G., Rossatto, D. R., Silva, L. C. R., Lau, O. M., … Franco, A. C. (2012). Ecological thresholds at the savanna-forest boundary: How plant traits, resources and fire govern the distributions of tropical biomes. *Ecology Letters*, *15*, 759–768.

Instituto Brasileiro de Geografia e Estatística (IBGE). (2012). *Manual técnico da vegetação brasileira (2a edição revista e ampliada)*. Rio de Janeiro, Brazil: IBGE.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., … Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, *19*, 1366–1379. https://doi.org/10.1111/ddi.12096

Lehmann, C. E. R., Anderson, T. M., Sankaran, M., Higgins, S. I., Archibald, S., Hoffmann, W. A., … Bond, W. J. (2014). Savanna vegetation-fire-climate relationships differ among continents. *Science*, *343*(6170), 548–552.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18–22.

Matteo, R. G., Felicísimo, A. M., Pottier, J., Guisan, A., & Muñoz, J. (2012). Do stacked species distribution models reflect altitudinal diversity patterns? *PLoS One*, *7*, e32586.

McKnight, M. W., White, P. S., McDonald, R. I., Lamoreux, J. F., Sechrest, W., Ridgely, R. S., & Stuart, S. N. (2007). Putting beta-diversity on the map: Broad-scale congruence and coincidence in the extremes. *PLoS Biology*, *5*, e272. https://doi.org/10.1371/journal.pbio.0050272

Melo, A. S., Rangel, T. F. L. V. B., & Diniz-Filho, J. A. F. (2009). Environmental drivers of beta-diversity patterns in New-World birds and mammals. *Ecography*, *32*, 226–236. https://doi.org/10.1111/j.1600-0587.2008.05502.x

Moulatet, G. M., Zuquim, G., Figueiredo, F. O. G., Lehtonen, S., Emilio, T., Ruokolainen, K., & Tuomisto, H. (2017). Using digital soil maps to infer edaphic affinities of plant species in Amazonia: Problems and prospects. *Ecology and Evolution*, *7*, 1–15.

Mucina, L. (2019). Biome: Evolution of a crucial ecological and biogeographical concept. *New Phytologist*, *222*, 97–114.

Oliveira-Filho, A. T. (2017). *NeoTropTree, Flora arbórea da região Neotropica: Um banco de dados envolvendo biogeographica, diversidade e conservação*. Universidade Federal de Minas Gerais. Retrieved from http://www.neotroptree.info/

Oliveira-Filho, A. T., & Fontes, M. A. (2000). Patterns of floristic differentiation among Atlantic Forests in Southeastern Brazil and the influence of climate. *Biotropica*, *32*, 793–810.

Pennington, R. T., Prado, D. E., & Pendry, C. (2000). Neotropical seasonally dry forests and Quaternary vegetation changes. *Journal of Biogeography*, *27*, 261–273. https://doi.org/10.1046/j.1365-2699.2000.00397.x

Peterson, A. T., Papeș, M., & Eaton, M. (2007). Transferability and model evaluation in ecological niche modeling: A comparison of GARP and MaxEnt. *Ecography*, *30*, 550–560. https://doi.org/10.1111/j.0906-7590.2007.05102.x

Pineda, E., & Lobo, J. M. (2009). Assessing the accuracy of species distribution models to predict amphibian species richness patterns. *Journal of Animal Ecology*, *78*, 182–190. https://doi.org/10.1111/j.1365-2656.2008.01471.x

Pottier, J., Dubuis, A., Pellissier, L., Maiorano, L., Rossier, L., Randin, C. F., … Guisan, A. (2013). The accuracy of plant assemblage prediction from species distribution models varies along environmental gradients. *Global Ecology and Biogeography*, *22*, 52–63. https://doi.org/10.1111/j.1466-8238.2012.00790.x

Purves, D., Scharlemann, J. P. W., Harfoot, M., Newbold, T., Tittensor, D. P., Hutton, J., & Emmot, S. (2013). Ecosystems: Time to model all life on Earth. *Nature*, *493*(7432), 295–297.

Queiroz, L. P., Cardoso, D., Fernandes, M. F., & Moro, M. F. (2017). Diversity and evolution of flowering plants of the Caatinga domain. In J. M. Cardoso da Silva, I. R. Lean, & M. Tabarelli (Eds.), *Caatinga: The largest tropical dry forest region in South America* (pp. 4–6). Berlin, Germany: Springer.

Scheffers, B. R., Joppa, L. N., Pimm, S. L., & Laurance, W. F. (2012). What we know and don't know about Earth's missing biodiversity. *Trends in Ecology and Evolution*, *27*, 501–510. https://doi.org/10.1016/j.tree.2012.05.008

Silva de Miranda, P., Oliveira-Filho, A. T., Pennington, R. T., Neves, D. M., Baker, R. T., & Dexter, K. G. (2018). Using tree species inventories to map biomes and assess their climatic overlaps in lowland tropical South America. *Global Ecology and Biogeography*, *27*, 899–912. https://doi.org/10.1111/geb.12749

Šímová, I., Violle, C., Svenning, J.-C., Kattge, J., Engemann, K., Sandel, B., … Enquist, B. J. (2018). Spatial patterns and climate relationships of major plant traits in the New World between woody and herbaceous species. *Journal of Biogeography*, *45*, 895–916.

Socolar, J. B., Gilroy, J. J., Kunin, W. E., & Edwards, D. P. (2016). How should beta-diversity inform biodiversity conservation? *Trends in Ecology and Evolution*, *31*, 67–80. https://doi.org/10.1016/j.tree.2015.11.005

Stolar, J., & Neilsen, S. E. (2015). Accounting for spatially biased sampling effort in presence-only species distribution modelling. *Diversity and Distributions*, *21*, 595–608. https://doi.org/10.1111/ddi.12279

VanDerWal, J., Murphy, H. T., Kutt, A. S., Perkins, G. C., Bateman, B. L., Perry, J. J., & Reside, A. E. (2013). Focus on poleward shifts in species' distribution underestimates the fingerprint of climate change. *Nature Climate Change*, *3*, 239–243. https://doi.org/10.1038/nclimate1688

Veenendaal, E. M., Torello-Raventos, M., Miranda, H. S., Sato, N. M., Oliveras, I., van Langevelde, F., … Lloyd, J. (2018). On the relationship between fire regime and vegetation structure in the tropics. *New Phytologist*, *218*, 153–166. https://doi.org/10.1111/nph.14940

Wan, Z. (2014). New refinements and validation of the collection-6 MODIS land-surface temperature/emissivity product. *Remote Sensing of Environment*, *140*, 36–45. https://doi.org/10.1016/j.rse.2013.08.027

Wan, Z., & Dozier, J. (1996). A generalized split-window algorithm for retrieving land-surface temperature from space. *IEEE Transactions on Geoscience and Remote Sensing*, *34*, 892–905. https://doi.org/10.1109/36.508406

Warren, R., VanDerWal, J., Price, J., Welbergen, J. A., Atkinson, I., Ramirez-Villegas, J., … Lowe, J. (2013). Quantifying the benefit of early climate change mitigation in avoiding biodiversity loss. *Nature Climate Change*, *3*, 678–682. https://doi.org/10.1038/nclimate1887

Whittaker, R. H. (1970). *Communities and ecosystems*. New York, NY: MacMillan.

Wiegand, T., & Moloney, K. A. (2013). *Handbook of spatial point-pattern analysis in ecology*. Boca Raton, FL: CRC Press.

Wilson, A. M., & Jetz, W. (2016). Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. *PLoS Biology*, *14*, e1002415. https://doi.org/10.1371/journal.pbio.1002415

Xu, W., Svenning, J.-C., Chen, G., Chen, B., Huang, J., & Ma, K. (2017). Plant geographical range size and climate stability in China: Growth form matters. *Global Ecology and Biogeography*, *27*, 506–517. https://doi.org/10.1111/geb.12710

Zhang, M.-G., Slik, J. W. F., & Ma, K.-P. (2016). Using species distribution modelling to delineate the botanical richness patterns and phytogeographical regions of China. *Scientific Reports*, *6*, 22400.

BIOSKETCHES

Peter W. Moonlight, **Tiina E. Särkinen, Domingos Cardoso** and **R. Toby Pennington** are part of the NERC-Newton-funded project including an interdisciplinary group of U.K. and Brazilian scientists aiming to understand the distribution and characteristics of the neglected seasonally dry tropical forests of the Caatinga region. **Domingos Cardoso** and **Gustavo Ramos** are funded by a Royal Society Advanced Fellowship. This paper is the result of a collaboration between these groups and **Pedro L. Silva de Miranda, Kyle G. Dexter** and **Ary Oliveira-Filho**, who work on the distribution of biomes across lowland tropical South America and the environmental drivers behind them.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Moonlight PW, Silva de Miranda PL, Cardoso D, et al. The strengths and weaknesses of species distribution models in biome delimitation. *Global Ecol Biogeogr*. 2020;00:1–15. https://doi.org/10.1111/geb.13149